JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer
ISSN: 2615-577X (Online)

Analisis *Clustering* Penyebaran *Corona Virus Disease* 2019 (Covid-19) di Indonesia Tahun 2021 Menggunakan Algoritma *K-Means*

Annisa Rachmawati[#1], Moch. Hafid T[#2], Dede Irmayanti[#3]

[#]*Teknik Informatika Sekolah Tinggi Teknologi Wastukancana*
Jl.Cikopak No.53 Sadang, Purwakarta, Jawa Barat

[1]annisarachmawati26@wastukancana.ac.id,[3]dedeirmayanti@wastukancana.ac.id

[2]mhafid@wastukancana.ac.id

Abstract— ***The problem that will be discussed for this research is the case of the spread of Covid-19 which is increasingly in the spotlight and disturbing in the world, especially in Indonesia. The purpose of this study was to analyze the clustering of the spread of Covid-19 in Indonesia in 2021. The number of datasets obtained was 11,741 data on the spread of Covid-19 from Januari to December 2021 from the kaggle site. This study uses data mining techniques using the k-means clustering algorithm to determine areas affected by the spread of Covid-19 in Indonesia in 2021. The algorithm used in the k-means algorithm, which is a partition-based method that uses representative objects called medoids as central points or centroids. In this study the authors use the KDD method in data pre-processing which includes the planning stage which consists of (problem identification, determining objectives, literature study), data collection, data selection, pre-pocessing stage (cleaning data), data transformation, data mining, and interpretation / evaluation. Data mining evaluation uses silhouette coefficient testing, this method is a cluster evaluation method that combines cohessian and separation methods. Cohessian is measured by calculating the average distance of each object in a cluster with the closest cluster. Based on the research and discussion of the results that have been carried out, the clusters of the spread of Covid-19 in Indonesia using the k-means clustering algorithm can be grouped into 4 clusters, namely cluster 0 as many ad 318 items, cluster 1 as many as 84 items, cluster 2 as many as 224 items, cluster 3 as many as 373 items.***

Keywords— *Covid-19, Kaggle, Data Mining,* KDD*, K-means clustering, Silhouette coefficient*

Abstrak— **Permasalahan yang akan dibahas untuk penelitian ini adalah kasus penyebaran Covid-19 yang semakin menjadi sorotan dan meresahkan di dunia khususnya di Indonesia. Tujuan dari penelitaian ini untuk menganalisis *clustering* penyebaran Covid-19 di Indonesia Tahun 2021. Jumlah dataset yang diperoleh sebanyak 11.741 data penyebaran Covid-19 dari bulan Januari sampai dengan bulan Desember Tahun 2021 dari situs *kaggle*. Penelitian ini menggunakan teknik *data mining clustering* untuk menentukan daerah yang terdampak penyebaran Covid-19 di Indonesia Tahun 2021. Algoritma yang digunakan adalah Algoritma *K-Means* yaitu metode berbasis partisi yang menggunakan objek representatif yang disebut *medoids* sebagai titik pusat atau *centroid*. Pada penelitian ini penulis menggunakan metode *Knowledge Discovery in Database* (KDD) dalam pengolahan data yang meliputi tahap perencanaan yang terdiri dari (identifikasi masalah, menentukan tujuan, studi pustaka), pengumpulan data, data *selection*, tahap *pre-processing* (*cleaning* data), transformasi data, *data mining*, dan interpretation/evaluasi. Evaluasi data mining menggunakan pengujian *silhouette coefficient*, metode ini merupakan metode evaluasi *cluster* yang menggabungkan metode *cohessian* dan *separation*. *Cohessian* yang diukur dengan menghitung jarak rata-rata setiap objek dalam sebuah *cluster* dengan *cluster* terdekatnya. Berdasarkan penelitian dan pembahasan hasil yang sudah dilakukan, maka *cluster* penyebaran Covid-19 di Indonesia menggunakan algoritma *k-means clustering* dapat dikelompokkan menjadi 4 *cluster* yaitu *cluster* 0 sebanyak 318 *items*, *cluster* 1 sebanyak 84 *items*, *cluster* 2 sebanyak 224 *items*, *cluster* 3 sebanyak 373 *items*.**

Kata kunci— Covid-19, *Kaggle, Data Mining,* KDD, *K-means clustering, Silhouette coefficient*

I. PENDAHULUAN

Covid-19 merupakan virus yang menyerang sistem pernapasan, memberi dampak buruk bagi kesehatan yang disertai dengan gejala yang ringan maupun yang berat. Virus ini, merupakan virus yang tidak diprediksi akan terjadi sebelumnya. Tanda dan gejala Covid-19 ini tergolong berat, yaitu terjadinya sindrom pernapasan akut, menyebabkan pneumonia, gagal ginjal, dan yang paling fatal berakibat kematian. Sedangkan gejala ringannya deman, bersin, sakit pada tenggorokan dan lain sebagainya [1].

Salah satu permasalahan yang akan dibahas dalam penelitian ini adalah kasus penyebaran Covid-19 yang semakin menjadi sorotan dan meresahkan di dunia khususnya di Indonesia. Penulis juga ingin mengetahui data klaster penyebaran covid-19 dari bulan Januari sampai dengan bulan Desember tahun 2021 di setiap wilayah yang ada di Indonesia agar dapat meminimalisir dampak yang dihasilkan.

Teknik *cluster* merupakan proses pengelompokan sekumpulan atau beberapa objek data ke dalam beberapa kelompok atau *cluster* sehingga objek dalam sebuah *cluster*

memiliki kemiripan yang tinggi, tapi sangat berbeda dengan objek dalam *cluster* lain [2].

Algoritma *k-means* merupakan metode berbasis partisi yang menggunakan objek representatif yang disebut *medoids* sebagai titik pusat atau *centroid*. Algoritma ini menerima masukan berupa data tanpa label kelas. Pada algoritma *k-means* komputer menerima data-data yang tidak diketahui kelasnya terlebih dahulu lalu mengelompokannya. Input yang diterima adalah data dan jumlah kelompok (*cluster*) yang diinginkan [2] .

Berdasarkan latar belakang yang telah diuraikan tersebut maka dari itu peneliti mengangkat tema yang berjudul: "Analisis *Clustering* Penyebaran *Corona Virus Disease* 2019 (Covid-19) di Indonesia Tahun 2021 Menggunakan Algoritma *K-Means"*.

II. TINJAUAN PUSTAKA

A. Penelitian Terdahulu

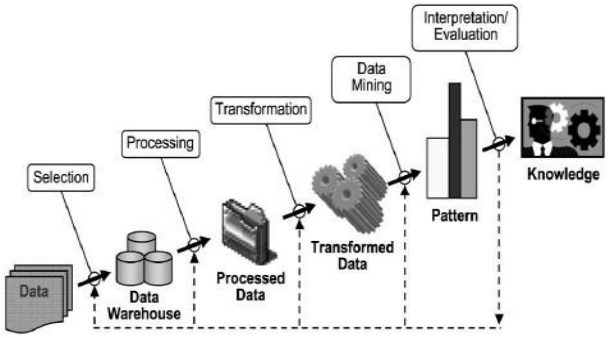
Solmin Paembonan dan Abduh Hisma melakukan penelitian untuk menerapkan metode *silhouette coefficient* untuk evaluasi *clustering* obat menggunakan algoritma k-means clustering. Dari hasil proses perhitungan *Silhoutte Coefficient* terhadap data obat maka hasil *Silhoutte Coefficient* yang maksimum adalah pada saat k = 2 dengan nilai *Silhoutte* = 0,4854. Nilai rata-rata *Silhoutte Coefficient* untuk semua cluster seperti yang terlihat pada gambar. Berdasarkan hasil pengujian emberikan informasi mengenai jumlah *cluste*r dan nilai *silhouette* yang maksimum dengan hasil *clustering* untuk k = 2 dan nilai *silhouette* = 0,4854, dan ketika jumlah *cluster* (k) semakin besar maka nilai *silhouette* yang dihasilkan cenderung lebih kecil dibanding jumlah *cluster* sebelumnya.

B. *Data Mining*

Data mining adalah proses menganalisa data dari perspektif yang berbeda dan menyimpulkannya menjadi informasi-informasi penting yang dapat dipakai untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya. Secara teknis, *data mining* disebut sebagai proses untuk menemukan korelasi atau pola dari ratusan atau ribuan *field* dari sebuah relasional *database* yang benar [3]. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. *Data mining* merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual [4].

C. *Knowledge Discovery in Database* (KDD)

KDD merupakan tahapan dalam penggalian data untuk mendapatkan informasi yang berharga. Adapun bentuk umum dari proses KDD dapat dijelaskan pada Gambar 1 berikut :



Gambar 1. Tahapan KDD [4].

D. Algoritma *K-Means*

Algoritma *K-Means* yang dimaksudkan sebagai konstanta jumlah *cluster* yang diinginkan, *Means* dalam hal ini berarti nilai suatu rata-rata dari suatu grup data yang dalam hal ini didefinisikan sebagai *cluster*, sehingga *K-Means Clustering* adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa *supervise* (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode *K-Means* berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama atau sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain [5].

Berikut adalah rumus untuk menentukan jarak data dari masing-masing centroid:

$$d\ (P,Q) = \sqrt{\sum_{j=1}^{P} (xj(P) - xj\big((Q)\big)2}$$

Rumus 1. Algoritma *K-Means* [5].

Keterangan :
d = Titik jarak
P = Data *record*
Q = Data *centroid*

E. *Silhouette Coefficient*

Metode ini merupakan metode evaluasi *cluster* yang menggabungkan metode *cohessian* dan *separation*. *Cohessian* diukur dengan menghitung jarak rata - rata setiap objek dalam sebuah *cluster* dengan *cluster* terdekatnya. Jarak antara data dengan menggunakan rumus *Euclidean distance* [6].

Berikut langkah-langkah untuk menghitung nilai *Silhouette Coefficient* :

1. Hitung rata – rata jarak dari suatu dokumen misalkan *i* dengan semua dokumen lain yang berada dalam satu cluster dengan *j* adalah dokumen lain dalam satu cluster *A* dan *d(i,j)* adalah jarak antara dokumen *i* dengan *j*.

$$\alpha(i) = \frac{1}{|A|-1} \sum j \in_{Aj \neq i} d(i,j)$$

Rumus 2. Jarak antara dokumen i dengan j [6].

2. Hitung rata-rata jarak dari dokumen *i* tersebut dengan semua dokumen di *cluster* lain, dan diambil nilai terkecilnya.

$$d(i,C) = \frac{1}{|A|} \sum j \in C \ \ d(i,j)$$

Rumus 3. Jarak rata-rata dokumen i dengan semua objek [6].

Dengan *d(i,C)* adalah jarak rata-rata dokumen *i* dengan semua objek pada *cluster* lain *C* dimana $A \neq C$. *b(i)* = min $C \neq A$ *d(i,C)*.

3. Nilai *Silhouette Coefficient* adalah :

$$S(i) = \frac{b(i) - \alpha(i)}{\max(a,i),b(i))}$$

Rumus 2. 1 Nilai *Silhoutte Coefficient* [6]

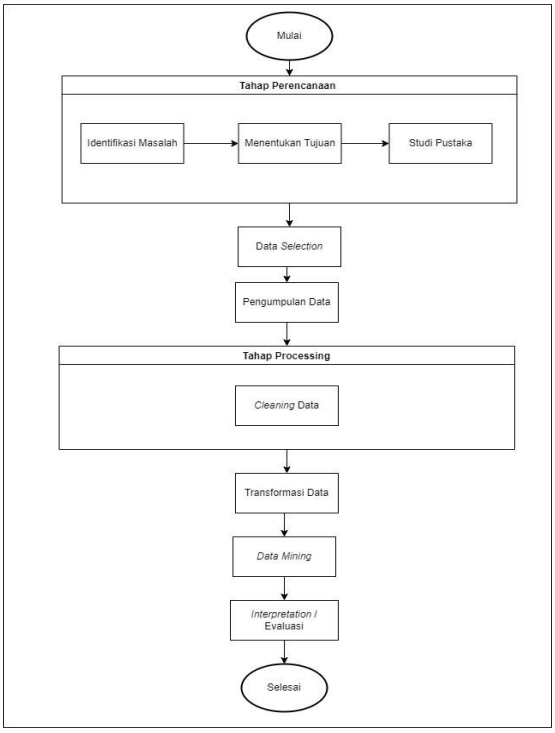
Hasil perhitungan *Silhouette Coefficient* memiliki *range* antara -1 hingga 1. Dikatakan baik apabila bernilai positif, hal ini berarti titik sudah berada di *cluster* yang tepat. Sedangkan jika nilainya negatif ini menandakan terjadinya *over lapping* sehingga titik berada di *cluster* yang tidak tepat. Jika nilainya 0, ini berarti berada di antara dua *cluster*. Berikut ini adalah representasi dari nilai *Silhoutte Coefficient*.

F. *Davies Bouldin Index* (DBI)

Davies Bouldin Index (DBI) adalah matriks untuk mengevaluasi atau mempertimbangkan hasil algoritma *clustering*. Pertama kali diperkenalkan oleh *David L. Davies* dan *Donald W. Bouldin* pada tahun 1979. Dengan menggunakan DBI suatu *cluster* akan dianggap memiliki skema *clustering* yang optimal adalah yang memiliki DBI minimal [7].

III. METODE PENELITIAN

Penelitian ini melakukan pengelompokan sebaran virus covid-19 di Indonesia Tahun 2021 menggunakan algoritma *K-Means Clustering* dengan tahapan-tahapan sebagai berikut :



Gambar 2. Kerangka Pemikiran [8].

1. Tahap Perencanaan
   Tahap perencanaan adalah tahapan-tahapan yang harus direncanakan oleh peneliti saat akan melakukan penelitian. Berikut ini penjelasan langkah-langkah dalam tahap perencanaan:
   a. Identifikasi Masalah
      Kegiatan ini dilakukan untuk mengelompokan data penyebaran Covid-19 di Indonesia Tahun 2021 menggunakan algoritma *k-means*.
   b. Menentukan Tujuan
      Tahap ini bertujuan untuk menganalisis clustering penyebaran *Corona Virus Disease* 2019 *(Covid-19)* di Indonesia Tahun 2021.
   c. Studi Pustaka / *Literature Review*
      Studi pustaka dalam penelitian ini didapatkan dari referensi jurnal dan buku yang ada di perpustakaan STT Wastukancana.
2. Pengumpulan Data
   Pada tahap *sampel* disini bertujuan sebagai pengumpulan data pada situs *kaggle* dengan mendapatkan data tanggal terjadinya penyebaran covid-19 dari bulan Januari 2021 sampai bulan Desember 2021.
3. Data *Selection*
   Pada tahap seleksi ini dari 29 atribut tidak semua atribut digunakan hanya beberapa atribut yang dibutuhkan untuk dapat diproses. Dipilihlah 13 atribut saja yang terdiri dari tanggal, kode lokasi, lokasi, kasus baru 1, kematian baru, sembuh, kasus baru 2, Total kasus, total kematian, total sembuh, total kasus aktif, *longitude*, dan *latitude*. Alasan dipilihnya hanya 13 atribut saja karena dari pemilihan data yang

relevan akan mudah diterima dari koleksi data yang ada.

4. Tahap *Pre-processing* Data
   a. *Cleaning* Data
      Cleaning data atau pembersihan data dilakukan untuk membersihkan data dengan cara melengkapi data, menghapus data duplikat dan data kosong serta data yang tidak digunakan. Data yang sudah pilih pada tahap seleksi data kemudian dibersihkan untuk menghilangkan *missing value* dan *redundant* data.
5. Transformasi Data
   Data ditransformasikan dalam bentuk yang sesuai untuk proses data mining. Dalam penelitian ini data total kasus dikelompokan menjadi 3 yaitu *suspek*, *probable*, dan konfirmasi. Data ini didapatkan dari hasil total rata-rata penilaian gugus tugas penanganan covid-19 di Indonesia Tahun 2021 dengan hasil yang signifikan.
6. *Data Mining*
7. Evaluasi
   Tahapan ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Evaluasi pada penelitian ini menggunakan *Silhouette Coefficient*. Evaluasi dilakukan untuk menghitung jarak rata-rata setiap objek dari sistem yang sudah dibuat berdasarkan hasil dari *clustering*.

IV. PENGOLAHAN DATA DAN PEMBAHASAN

4.1 Pengumpulan Data

Pada tahap pengumpulan data didapatkan data sebanyak 11.741 data kasus penyebaran covid-19 di Indonesia Tahun 2021 dengan 29 atribut.

4.2 Data *Selection*

Pada tahap seleksi ini dari 29 atribut tidak semua atribut digunakan hanya beberapa atribut yang dibutuhkan untuk dapat diproses. Dipilihlah 13 atribut saja yang terdiri dari tanggal, kode lokasi, lokasi, kasus baru 1, kematian baru, sembuh, kasus baru 2, Total kasus, total kematian, total sembuh, total kasus aktif, *longitude*, dan *latitude.* Alasan dipilihnya hanya 13 atribut saja karena dari pemilihan data yang relevan akan mudah diterima dari koleksi data yang ada.

4.3 *Pre-processing* Data

1. *Cleaning* Data

Cleaning data atau pembersihan data dilakukan untuk membersihkan data, dengan cara melengkapi data, menghapus data duplikat dan data kosong serta data yang tidak digunakan. Data yang sudah dipilih pada tahap seleksi data kemudian dibersihkan untuk menghilangkan *missing value* dan *redundant* data. Pada proses *cleaning* data menggunakan *Rapidminer*.

Pada tahap pembersihan data tidak terdapat data yang nilainya *null* atau *Missing*, sehingga dapat melanjutkan proses selanjutnya. Setelah hasil dari proses *cleaning* data, maka jumlah data yang dihasilkan tetap menjadi 11.741 data. Selanjutnya dari 13 atribut dalam pemilihan data tersebut dapat diintegrasikan dan diambil 5 atribut utama untuk digabungkan menjadi satu tabel utama yang digunakan sebagai data akhir atau dataset untuk klasterisasi. Dalam hal ini proses data seleksi akan menghilangkan atribut tanggal, kode lokasi, lokasi, kasus baru 1, kematian baru, sembuh, kasus baru 2, dan total kasus aktif. Dari 5 atribut tersebut terdiri dari :

1. Total kasus
2. Total kematian
3. Total sembuh
4. Longitude
5. Latitude

Alasan dipilihnya hanya 5 atribut ini, karena dengan data Total kita akan mengetahui hasil akhir pengelompokan *cluster* dari penelitian ini sehingga dapat meminimalisir dampak yang dihasilkan.

4.4 Transformasi Data

Data ditransformasikan dalam bentuk yang sesuai untuk proses *data mining*. Dalam penelitian ini data Total kasus dikelompokan menjadi 3 yaitu *Suspek*, *Probable*, dan Konfirmasi. *Suspek* adalah seseorang yang menderita infeksi saluran pernafasan dengan gejala berat dan perlu menjalani perawatan di rumah sakit tanpa penyebab yang spesifik. *Probable* adalah orang yang masih dalam kategori suspek dan memiliki gejala berat dan gagal nafas atau meninggal dunia. Sedangkan kasus konfirmasi adalah orang yang sudah dinyatakan positif terinfeksi vius Corona berdasarkan hasil pemeriksaan. Menurut [9] untuk nilai status suspek berkisar ≤ 10.000 orang, probable ≥ 10.000 orang, sedangkan yang terkonfirmasi berkisar ≥ 100000 orang. Data ini didapatkan dari hasil total rata-rata penilaian gugus tugas penanganan covid-19 di Indonesia Tahun 2021 dengan hasil yang signifikan. Transformasi data atribut Total Kasus dapat dilihat pada Tabel 1 berikut**.**

Tabel 1**.** Transformasi Atribut Total Kasus

Total Kasus	Status Kasus
8753	Suspek
17694	Probable
18441	Probable
3671	Suspek
185690	Konfirmasi
12388	Probable
3866	Suspek
751270	Konfirmasi
3263	Suspek
91592	Probable
83446	Probable
85272	Konfirmasi
3136	Suspek
15402	Probable

Korespondensi: annisarachmawati26@wastukancana.ac.id

Setelah dikelompokkan, kemudian atribut total kasus di transformasikan menjadi numerik. Transformasi data total kasus menjadi numerik dapat dilihat pada Tabel 2 berikut.

Tabel 2. Transformasi Atribut Total Kasus Menjadi Numerik

Status Kasus	Inisialisasi
Suspek	1
Probable	2
Konfirmasi	3

Data Total kematian dikelompokan menjadi 3 yaitu Tidak Parah, Parah, dan Sangat Parah. Menurut (Kemenkes RI) bahwa untuk kategori Tidak Parah berkisar 20% atau bisa dikatakan ≤100 orang, untuk kategori Parah berkisar 54% atau ≥100 orang, sedangkan kategori sangat parah megalami 84% peningkatan kematian atau ≥1000 orang di tahun 2021. Transformasi data atribut total kematian dapat dilihat pada Tabel 3 berikut.

Tabel 3. Transformasi Atribut Total Kematian

Total Kematian	Status Kematian
358	Parah
517	Parah
537	Parah
Total Kematian	Status Kematian
117	Parah
3290	Sangat Parah
272	Parah
104	Parah
22329	Sangat Parah
56	Tidak Parah
1461	Sangat Parah
3631	Sangat Parah
6463	Sangat Parah
27	Tidak Parah
584	Parah

Setelah dikelompokan, kemudian atribut total kematian di transformasikan menjadi numerik. Transformasi data total kematian menjadi numerik dapat dilihat pada Tabel 4 berikut.

Tabel 4. Transformasi Atribut Total Kematian Menjadi Numerik

Status Kematian	Inisialisasi
Tidak Parah	1
Parah	2
Sangat Parah	3

Pada transformasi atribut total sembuh dikelompokan menjadi 1 yaitu Negatif. Nilai rata-rata ini sudah dikatakan valid oleh gugus tugas penanganan Covid-19 pada tahun 2021 bahwa untuk status sembuh tidak ada nilai minimum atau maksimum. Transformasi data atribut total sembuh dapat dilihat pada Tabel 5 berikut.

Tabel 5. Transformasi Atribut Total Sembuh

Total Sembuh	Status Sembuh
7149	Negatif
16223	Negatif
15844	Negatif
2653	Negatif
166349	Negatif
8289	Negatif
3327	Negatif
617936	Negatif
2352	Negatif
75883	Negatif
57043	Negatif
75691	Negatif
2804	Negatif
13861	Negatif

Tabel 6. Transformasi Atribut Total Sembuh Menjadi Numerik

Status Sembuh	Inisialisasi
Negatif	1

Dari hasil keseluruhan transformasi data di atas berikut inilah yang akan digunakan dalam proses perhitungan data mining menggunakan algoritma *K-Means Clustering.* Berikut ini Tabel 7 Hasil Transformasi Keseluruhan.

Tabel 7. Hasil Transformasi Keseluruhan

Total Kasus	Total Kematian	Total Sembuh	Longitude	Latitude
1	2	1	96,9	4,2
2	2	1	115,1	-8,4
2	2	1	106,1	-6,5
1	2	1	102,3	-3,5
3	3	1	106,8	-6,2
2	2	1	110,4	-7,9
1	2	1	122,4	0,7
3	3	1	113,9	-0,8
1	1	1	102,7	-1,7
2	3	1	107,6	6,9
2	3	1	110,2	-7,3

4.5 *Data Mining*

Berikut adalah proses data mining untuk mengelompokan suatu data:

1. Menentukan jumlah *cluster* dan menentukan koordinat titik tengah *cluster*. Kelompok *cluster* yang dibuat adalah 4 kelompok agar terlihat jarak antara titik *centroid*nya. Total jumlah inisiasi data dari setiap atribut tersebut untuk menentukan kelompok diambil dari penyebaran tingkat rendah, sedang, dan tinggi secara acak dan hasilnya seperti pada tabel 4.8 berikut :

Korespondensi: annisarachmawati26@wastukancana.ac.id

Tabel 8 Titik *Centroid* Awal

Cluster	Total Kasus	Total Kematian	Total Sembuh	Lon	Lat
CLUSTER 1	2	2	1	115,1	-8,4
CLUSTER 2	3	3	1	106,8	-6,2
CLUSTER 3	3	3	1	113,9	-0,8
CLUSTER 4	2	3	1	112,7	-7,7

2. Penentuan nilai *cluster* untuk dijadikan acuan dalam melakukan perhitungan jarak objek ke *centroid,* perhitungan jarak mengacu pada rumus *Euclidean Distance* dibawah ini:

$$d\ (P, Q) = \sqrt{\sum_{j=1}^{P} \big(xj\ (P) - xj(Q)\big)2}$$

Sumber: [5].

Keterangan :
d = data titik jarak (*Euclidean*)
P = data *record*
Q = data *centroid*

Rumus *Euclidean* merupakan perhitungan jarak antara centroid, perhitungan ini dilakukan di excel karena jumlah data yang sangat banyak.

3. Setelah jarak antara *centroid* dihitung dengan menggunakan rumus *Euclidean distance*, maka dilakukan pengelompokan *centroid* sesuai dengan hasil dari jarak antara *centroid* tersebut. Kemudian hasil dari perhitungan jarak tersebut digunakan untuk menentukan kelompok *clustering*. Penentuan dalam pengelompokan *centroid* adalah sebagai berikut.
   a. Jika jarak *centroid* 0 lebih kecil dari jarak *centroid* 1, *centroid* 2, dan *centroid* 3, maka termasuk kelompok *centroid* 0.
   b. Jika jarak *centroid* 1 lebih kecil dari jarak *centroid* 0, *centroid* 2, dan *centroid* 3, maka termasuk kelompok *centroid* 1.
   c. Jika jarak *centroid* 2 lebih kecil dari jarak *centroid* 0, *centroid* 1, dan *centroid* 3, maka termasuk kelompok *centroid* 2.
   d. Jika jarak *centroid* 3 lebih kecil dari jarak *centroid* 0, *centroid* 1, dan *centroid* 2, maka termasuk kelompok *centroid* 3.

Berikut adalah hasil perhitungan jarak *centroid* masing-masing dalam *cluster:*

Tabel 9. Jarak hasil perhitungan antara *Centroid* pada Iterasi 1

Hasil	C0	C1	C2	C3
1	491.000	209,170	317.000	393,250
2	0	75,730	61,200	7,250
3	4,100	2,580	95,330	46.000
4	188,850	30,540	144.850	127,800
5	75,730	0	79.570	38,060
6	22,340	17.850	64.660	5,330
7	137,100	293.970	77.500	166,650
8	61,200	85.570	0	50,050
9	63,700	42,060	132,250	141.000
10	291,340	173,250	99,980	239,170
11	3,770	12,560	15,900	6,410
12	7,250	4,690	50,050	0
13	86,890	61,700	14,330	65,320
14	29,250	86,200	9,090	30,380

Pada tabel 9 ini adalah hasil perhitungan jarak dengan masing-masing *cluster*. Jarak yang terlihat dengan masing-masing *centroid* di setiap *cluster*, jika jarak antara dua titik semakin dekat, maka akan semakin dekatlah pula kesamaan antara kedua titik tersebut. Data pertama mempunyai jarak paling dekat dengan *centroid* di kelompok C1 dibandingkan dengan *centroid* dari kelompok lain, maka bisa disimpulkan bahwa data pertama mempunyai karakteristik paling dekat, sehingga data pertama dapat dimasukan ke dalam kelompok C1. Pengelompokan *centroid* lainnya seperti pada Tabel 10 berikut.

Tabel 10. Pengelompokan *Centroid*

Hasil	C0	C1	C2	C3
1		*		
2	*			
3		*		
4		*		
5		*		
6				*
7			*	
8			*	
9		*		
10			*	
11	*			
12				*
13			*	
14			*	

4.5.1 Pengujian *Software Rapidminer*

Data akan dianalisis berdasarkan lokasi yang paling banyak terjadi penyebaran *virus corona*. Data dibuat dalam set baru dengan format .xlsx (*excel*) untuk bisa dianalisis. *Software* yang dipakai adalah *RapidMiner Studio*.

Setelah menghubungkan *operator*, langkah selanjutnya memilih jumlah *cluster* pada menu *parameters*, pada penelitian ini penulis memilih 4 *cluster*.

Gambar 3. Menu *Parameters*

Setelah selesai memilih jumlah cluster, kemudian pada bagian *measure types* pilih *NumericalMeasures* dan E*uclideanDistance* dengan optimasi sebanyak 9 kali karena jumlah nilai cluster berhenti pada cluster 9. Setelah selesai pada tahap pemilihan jumlah *cluster*, *measure types* dan *numerical measure* langkah berikutnya adalah menghubungkan kedua *operator* dan hubungkan juga pada proses *output* untuk mendapatkan hasil *run operator*nya.

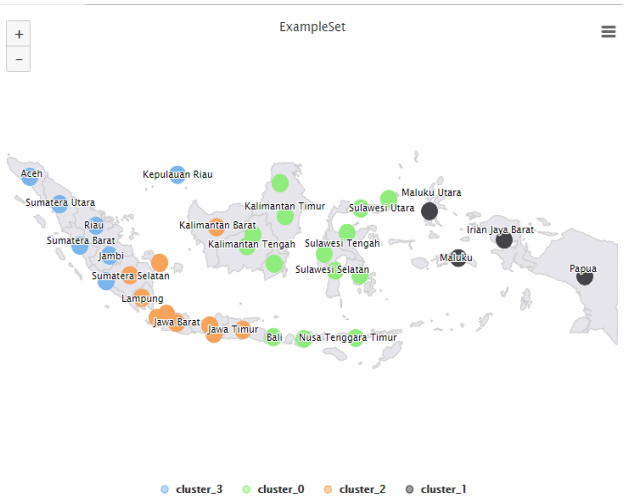
Tabel 11. Cluster Model

Cluster 0	318 Items
Cluster 1	84 Items
Cluster 2	224 Items
Cluster 3	373 Items
Total Number Of Items	999

Pada Tabel 11 yaitu *cluster model*. Tahapan ini bisa diketahui dari hasil komputasi menghasilkan beberapa *cluster model*, diantaranya untuk *cluster* 0 sebanyak 318 items, *cluster* 1 sebanyak 84 items, *cluster* 2 sebanyak 224 items, *cluster* 3 sebanyak 373 items. Hasil *cluster* tersebut dikelompokan menggunakan *RapidMiner* berdasarkan Total Kasus, Total Sembuh, Total Kematian, *Longitude*, *Latitude*. *Cluster model* pada *Rapidminer* ini berfungsi untuk menampilkan hasil yang lebih mendetail dan spesifik terkait dengan setiap *cluster*.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Total Kematian	1.539	1.523	1.578	1.530
Total Sembuh	1	1	1	1
Longitude	118.485	132.239	108.093	101.669
Latitude	-2.626	-2.447	-5.320	0.682

Gambar 4. *Centroid Cluster*



Gambar 5. Peta Sebaran Covid-19 Tahun 2021

Visualisasi hasil *cluster* menurut total kasus menggunakan aplikasi *RapidMiner* terhadap penyebaran covid 19 di Indonesia Tahun 2021 berdasarkan Total Kasus dengan menggunakan algoritma *K-Means Clustering*. Untuk pengelompokan *Cluster* 0 terdiri dari daerah Nusa Tenggara Timur, Bali, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Utara, Kalimantan Tengah, dan Kalimantan Timur. Pengelompokan *Cluster* 1 terdiri dari daerah Papua, Maluku, Maluku Utara, dan Irian Jaya Barat. Pengelompokan *Cluster* 2 terdiri dari daerah Jawa Barat, Jawa Timur, Lampung, dan Kalimantan Barat. Pengelompokan *Cluster* 3 terdiri dari daerah Kepulauan Riau, Sumatera Selatan, Sumatera Barat, Sumatera Utara, Jambi, Riau, dan Aceh. Visualisasi hasil *cluster* menurut total sembuh menggunakan aplikasi *RapidMiner* terhadap penyebaran covid 19 di Indonesia Tahun 2021 berdasarkan Total Sembuh dengan menggunakan algoritma *K-Means Clustering*. Untuk pengelompokan *Cluster* 0 terdiri dari daerah Nusa Tenggara Timur, Bali, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Utara, Kalimantan Tengah, dan Kalimantan Timur. Pengelompokan *Cluster* 1 terdiri dari daerah Papua, Maluku, Maluku Utara, dan Irian Jaya Barat. Pengelompokan *Cluster* 2 terdiri dari daerah Jawa Barat, Jawa Timur, Lampung, dan Kalimantan Barat. Pengelompokan *Cluster* 3 terdiri dari daerah Kepulauan Riau, Sumatera Selatan, Sumatera Barat, Sumatera Utara, Jambi, Riau, Sulawesi Utara dan Aceh. Visualisasi hasil *cluster* menurut total kematian menggunakan aplikasi *RapidMiner* terhadap penyebaran covid 19 di Indonesia Tahun 2021 berdasarkan Total Kematian dengan menggunakan algoritma *K-Means Clustering*. Untuk pengelompokan *Cluster* 0 terdiri dari daerah Sulawesi Barat, Kalimantan Selatan, Riau, Maluku Utara, Kalimantan Barat, Aceh, Sulawesi Tengah, Bali, Sulawesi Utara, Sulawesi Selatan, Kalimantan Timur, Kalimantan Tengah. Pengelompokan *Cluster* 1 terdiri dari daerah Maluku, Lampung, NTT, Maluku Utara. Pengelompokan *Cluster* 2 terdiri dari daerah Jawa Barat, Jawa Timur, Bali, Sumatera Barat, Kepulauan Riau, Kalimantan Utara, Kalimantan Barat, Lampung, Sumatera Selatan.

Pengelompokan C*luster* 3 terdiri dari daerah Sulawesi Utara, Sumatera Selatan, Papua, Riau, Kepulauan Riau, Sumatera Barat, Aceh, dan Sumatera Utara.

5 Evaluasi

Pada tahap ini peneliti melakukan analisis evaluasi perhitungan jarak terhadap nilai *silhouette coefficient* pada algoritma *k-means* dengan perhitungan jarak terhadap *centroid* dengan metode perhitungan yaitu *Euclidean distance*, *jaccard*, *cosine distance* serta menghitung nilai *silhouette coefficient* untuk setiap metode perhitungan jarak tersebut [10]. Berikut adalah perhitungan *Silhouette Coefficient* dengan menggunakan perhitungan manual dan *Microsoft Excel* dengan menggunakan *sampel* sebanyak 14 *sampel*.

Cluster 0 = {0}, {3,77}
Cluster 1 = {209,17}, {2,58}, {30,54}, {0}, {42,06}
Cluster 2 = {77,50}, {0}, {99,98}, {14,33}, {9,09}
Cluster 3 = {5,33}, {0}

1. *Cluster* 0
   a. Menghitung rata-rata objek dengan semua objek lain yang berada di dalam satu *cluster*:

$$\text{Cluster } 0 = \{0\} \rightarrow \alpha(i) = \sqrt{(1-0)\text{^}2} = 1$$
$$\{3{,}77\} \rightarrow \quad \alpha(i) \quad = \sqrt{(1-3{,}77)\text{^}2} = 2{,}77$$
$$= 3{,}77 / 2 = 1{,}885$$

   b. Menghitung rata-rata objek dengan semua objek lain yang berada pada *cluster* lain. Menghitung rata – rata dengan objek yang berada di *cluster* 1 dan menghitung rata-rata semua jarak pada *cluster* 1.

Tabel 12. Rata-rata Jarak *Cluster* 1

Jarak *Cluster* 1	Rata-rata
$C0_1 \rightarrow C1_1 \rightarrow \{0\} \rightarrow \{209{,}17$	209,17
$C0_1 \rightarrow C1_2 \rightarrow \{0\} \rightarrow \{2{,}58\}$	2,58
$C0_1 \rightarrow C1_3 \rightarrow \{0\} \rightarrow \{30{,}54\}$	30,54
$C0_1 \rightarrow C1_4 \rightarrow \{0\} \rightarrow \{0\}$	0
$C0_1 \rightarrow C1_5 \rightarrow \{0\} \rightarrow \{42{,}06\}$	42,06
$C0_2 \rightarrow C1_1 \rightarrow \{3{,}77\} \rightarrow \{209{,}17\}$	205,40
$C0_2 \rightarrow C1_2 \rightarrow \{3{,}77\} \rightarrow \{2{,}58)$	1,19
$C0_2 \rightarrow C1_3 \rightarrow \{3{,}77\} \rightarrow \{30{,}54\}$	26,77
$C0_2 \rightarrow C1_4 \rightarrow \{3{,}77\} \rightarrow \{0\}$	3,77
$C0_2 \rightarrow C1_5 \rightarrow \{3{,}77\} \rightarrow \{42{,}06\}$	38,29
Total	55,977

   c. Menghitung rata-rata objek dengan objek yang berada di *cluster* 2 dan menghitung rata-rata semua jarak pada *cluster* 2.

Tabel 13. Rata-rata Jarak *Cluster* 2

Jarak *Cluster* 2	Rata-rata
$C0_1 \rightarrow C2_1 \rightarrow \{0\} \rightarrow \{77{,}50\}$	77,50
$C0_1 \rightarrow C2_2 \rightarrow \{0\} \rightarrow \{0\}$	0
$C0_1 \rightarrow C2_3 \rightarrow \{0\} \rightarrow \{99{,}98\}$	99,98
$C0_1 \rightarrow C2_4 \rightarrow \{0\} \rightarrow \{14{,}33\}$	14,33
$C0_1 \rightarrow C2_5 \rightarrow \{0\} \rightarrow \{9{,}09\}$	9,09
$C0_2 \rightarrow C2_1 \rightarrow \{3{,}77\} \rightarrow \{77{,}50\}$	73,73
$C0_2 \rightarrow C2_2 \rightarrow \{3{,}77\} \rightarrow \{0\}$	3,77
$C0_2 \rightarrow C2_3 \rightarrow \{3{,}77\} \rightarrow \{99{,}98\}$	96,21
$C0_2 \rightarrow C2_4 \rightarrow \{3{,}77\} \rightarrow \{14{,}33\}$	10,56
$C0_2 \rightarrow C2_5 \rightarrow \{3{,}77\} \rightarrow \{9{,}09\}$	5,32
Total	39,049

   d. Menghitung rata-rata objek dengan objek yang berada di *cluster* 3 dan menghitung rata-rata semua jarak pada *cluster* 3.

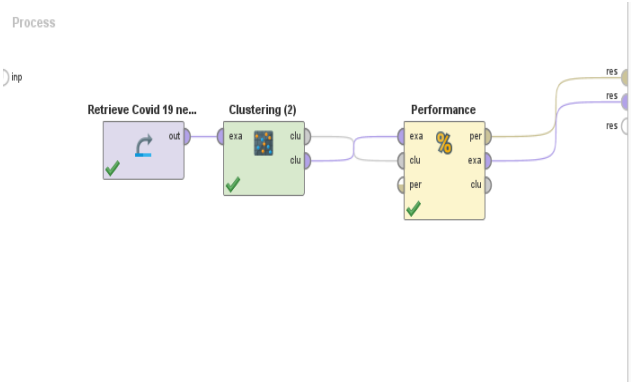
Tabel 14. Rata-rata Jarak *Cluster* 3

Jarak *Cluster* 3	Rata-rata
$C0_1 \rightarrow C3_1 \rightarrow \{0\} \rightarrow \{5{,}33\}$	5,33
$C0_1 \rightarrow C3_2 \rightarrow \{0\} \rightarrow \{0\}$	0
$C0_2 \rightarrow C3_1 \rightarrow \{3{,}77\} \rightarrow \{5{,}33\}$	1,56
$C0_2 \rightarrow C3_2 \rightarrow \{3{,}77\} \rightarrow \{0\}$	3,77
Total	2,665

   e. Nilai minimum dari rata-rata jarak *cluster* 1 dan *cluster* 3.
   b (*i*) = (2,665 < 55,977) = 2,665
   Maksimum = 55,977

   f. Menghitung nilai *silhouette coefficient*
   S (*i*) = (a*i* - b*i*) / max (a(*i*),b(*i*))= (1,885-2,665) / 55,977 = 0,0139343

Untuk perhitungan selanjutnya *cluster* 1 dan *cluster* 2 sama seperti perhitungan di atas. *Cluster* yang diuji hanya pada *cluster* 0, *cluster* 1, dan *cluster* 2 saja, karena *cluster* 3 menghasilkan data atau hasil yang sama. Dari hasil pengujian *Silhouette Coefficient* di atas dapat disimpulkan bahwa semakin nilai *silhouette coefficient* mendekati nilai 1, maka semakin baik pengelompokan data dalam satu *cluster*.



Gambar 6. Proses *Performance Vector*

Tabel 15. Hasil Proses *Performance Vector*

Avg. *within centroid distance*	-5.033

Avg. *within centroid distance_cluster*_0	-6.254
Avg. *within centroid distance_cluster*_1	-5.347
Avg. *within centroid distance_cluster*_2	-3.465
Avg. *within centroid distance_cluster*_3	-4.688
Davies Bouldin	-0.189

Pada Tabel 15 Merupakan data *performance* dari hasil *Cluster K-Means data mining*.

Pengujian ini dilakukan untuk mencari *Silhouette Coefficient* dengan hasil *performance* pada *RapidMiner*.

Semakin kecil nilai *Davies Bouldin Index* yang diperoleh (non-negatif >=0) maka semakin baik *cluster* yang diperoleh dari pengelompokan menggunakan metode *clustering* [7]. Hasil perhitungan menggunakan algoritma *K-Means* menunjukan nilai -0.189 yang artinya tidak cukup baik atau kurang representatif karena dekatnya perhitungan yang relatif rendah.

V. KESIMPULAN

Berdasarkan penelitian dan pembahasan hasil yang sudah dilakukan, maka kesimpulan yang dapat diperoleh dari penelitian ini yaitu Algoritma *K-Means* dalam penelitian ini berhasil diterapkan untuk menentukan *cluster* data penyebaran *virus corona* di Indonesia. Analisis ini berhasil mengelompokan data penyebaran *virus corona* di Indonesia. Metode *K-Means* ini mempunyai ketelitian yang tinggi dalam mengukur sebuah data sehingga dapat menghasilkan *cluster* 0 sebanyak 318 *items*, *cluster* 1 sebanyak 84 *items*, *cluster* 2 sebanyak 224 *items*, dan *cluster* 3 sebanyak 373 *items*. Dari hasil visualisasi *Cluster* menggunakan aplikasi *Rapidminer* menunjukan bahwa *Cluster* 0 ada pada daerah Nusa Tenggara Timur, Bali, Sulawesi Selatan, Sulawesi Tengah, Sulawesi Utara, Kalimantan Tengah, dan Kalimantan Timur, untuk pengelompokan *Cluster* 1 terdiri dari daerah Papua, Maluku, Maluku Utara, dan Irian Jaya Barat. *Cluster* 2 terdiri dari daerah Jawa Barat, Jawa Timur, Lampung, dan Kalimantan Barat, *Cluster* 3 terdiri dari daerah Kepulauan Riau, Sumatera Selatan, Sumatera Barat, Sumatera Utara, Jambi, Riau, dan Aceh. Hasil *performance* dari proses *clustering* menggunakan algoritma *K-Means* adalah -0.189 yang artinya tidak cukup baik. Nilai rata-rata s(*i*) yang dievaluasi dari total 14 *sampel* menggunakan *silhouette coefficient* yaitu *cluster* 0 menghasilkan 0,0139343 yang artinya baik, *cluster* 1 menghasilkan -0,00693 yang artinya tidak cukup baik, *cluster* 2 menghasilkan 0,958361 yang artinya baik. Sedangkan *cluster* 3 menghasilkan data atau hasil yang sama dengan *cluster* lain.

REFERENSI

[1] Riani, "Latar belakang covid 19," 2021.

[2] N. Dwitri and Tampubolon, "Penerapan Algoritma K-means Dalam Menentukan Tingkat Penyebaran Pandemi COVID-19 Di Indonesia," vol. 4, no. 1, pp. 128–132, 2020.

[3] D. N. Sari and I. Yunita, "Tingkat Keparahan Dan Risiko Penyebaran COVID-19 Di Indonesia Dengan Menggunakan K-(Severity Level and Spreading Risk of Covid-19 in Indonesia by K-Means Clustering)," pp. 210–216, 2020.

[4] C. Nas, "Data Mining Pengelompokan Bidang Keahlian Mahasiswa Menggunakan Algoritma K-Means (Studi Kasus : Universitas Cic Cirebon)," vol. 09, no. 1, pp. 1–14, 2020.

[5] W. Lestari, "Clustering Data Mahasiswa Menggunakan Algoritma K-Means Untuk Menunjang Strategi Promosi (Studi Kasus : STMIK Bina Bangsa Kendari)," vol. 4, no. 2, pp. 35–48, 2019.

[6] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coeficient Untuk Evaluasi Clutering Obat," vol. 6, no. 2, pp. 48–54, 2021.

[7] N. S. Sufajar Butsianto1, "Penerapan Data Mining Terhadap Minat Siswa Dalam Mata Pelajaran Matematika Dengan Metode K-Means," vol. 10, pp. 113–121, 2019.

[8] H. Jaya, "Konsep Data Mining KDD," vol. 01, 2020.

[9] H. Supriyadi, "Kontak Erat COVID-19 Pada Anggota POGI Muda Prevention of Suspect , Probable , Confirmation , and Close Contact of COVID-19 in Young POGI Members," vol. 9, no. 1, pp. 114–118, 2021.

[10] R. Hidayati and Zubair, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," vol. 20, no. 2, pp. 186–197, 2021.

Korespondensi: annisarachmawati26@wastukancana.ac.id